HHS Public Access

Author manuscript

Lancet Oncol. Author manuscript; available in PMC 2021 June 28.

Published in final edited form as:

Lancet Oncol. 2019 July ; 20(7): 938–947. doi:10.1016/S1470-2045(19)30333-X.

Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study

Philipp Tschandl, **Noel Codella**, **Bengü Nisa Akay**, **Giuseppe Argenziano**, **Ralph P Braun**, **Horacio Cabo**, **David Gutman**, **Allan Halpern**, **Brian Helba**, **Rainer Hofmann-Wellenhof**, **Aimilios Lallas**, **Jan Lapins**, **Caterina Longo**, **Josep Malvehy**, **Michael A Marchetti**, **Ashfaq Marghoob**, **Scott Menzies**, **Amanda Oakley**, **John Paoli**, **Susana Puig**, **Christoph Rinner**, **Cliff Rosendahl**, **Alon Scope**, **Christoph Sinz**, **H Peter Soyer**, **Luc Thomas**, **Iris Zalaudek**, **Harald Kittler**

ViDIR Group, Department of Dermatology (P Tschandl PhD, C Sinz MD, H Kittler MD) and Center for Medical Statistics, Informatics and Intelligent Systems (CeMSIIS) (C Rinner PhD), Medical University of Vienna, Vienna, Austria; IBM Research AI, T J Watson Research Center, Yorktown Heights, NY, USA (N Codella PhD); Department of Dermatology, Medicine Faculty, Ankara University, Ankara, Turkey (B N Akay MD); Dermatology Unit, University of Campania, Naples, Italy (Prof G Argenziano PhD); Skin Cancer Center, Department of Dermatology, University Hospital Zürich, Zürich, Switzerland (R P Braun MD); Department of Dermatology, Instituto de Investigaciones Médicas, Buenos Aires, Argentina (Prof H Cabo MD); Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA (D Gutman PhD); Dermatology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA (Prof A Halpern MD, M A Marchetti MD); Kitware, Clifton Park, NY, USA (B Helba BS); Department of Dermatology, Medical University Graz, Graz, Austria (R Hofmann-Wellenhof MD); First Department of Dermatology, Aristotle University, Thessaloniki, Greece (A Lallas PhD); Department of Dermatology, Karolinska University Hospital and Karolinska Institutet, Stockholm, Sweden (J Lapins MD); Department of Dermatology, University of Modena and Reggio Emilia, Modena, Italy (C Longo PhD); Azienda Unità Sanitaria Locale—IRCCS di Reggio Emilia, Centro Oncologico ad Alta Tecnologia Diagnostica-Dermatologia, Reggio Emilia, Italy (C Longo); Melanoma Unit, Dermatology Department, Hospital Clínic Barcelona, Universitat de Barcelona, IDIBAPS, Barcelona, Spain (J Malvehy MD, S Puig MD); Centro de Investigación Biomédica en Red de Enfermedades Rarasd (CIBER ER), Instituto de Salud Carlos III, Barcelona, Spain (J Malvehy, S Puig); Memorial Sloan Kettering Cancer Center, Hauppauge, NY, USA (A Marghoob MD); Sydney Melanoma Diagnostic Centre & Sydney Medical School, Faculty of Medicine and

Correspondence to: Dr Harald Kittler, Department of Dermatology, Medical University of Vienna, 1090 Vienna, Austria harald.kittler@meduniwien.ac.at.

Contributors

PT and HK came up with the concept and designed the study, drafted the manuscript, did the statistical analysis, had full access to all the data in the study, take responsibility for the integrity of the data and the accuracy of the data analysis, and critically reviewed the manuscript for important intellectual content. PT, NC, and HK drafted the response to the reviewers. PT and CS created the study figures. HK and AH provided overall supervision to the study. All authors contributed to data collection, data analysis, or data interpretation, and provided administrative, technical, or material support.

See Online for appendix

For the **ISIC 2018 challenge website** see https://challenge2018.isic-archive.com/

Health, The University of Sydney, Sydney, NSW, Australia (Prof S Menzies MD); Department of Dermatology, Waikato District Health Board and Waikato Clinical Campus, University of Auckland, Hamilton, New Zealand (A Oakley MBChB); Department of Dermatology and Venereology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden (J Paoli MD); School of Clinical Medicine, University of Queensland (C Rosendahl PhD) and Dermatology Research Centre, The University of Queensland Diamantina Institute (Prof H P Soyer MD), University of Queensland, Brisbane, QLD, Australia; Medical Screening Institute, Sheba Medical Center and Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel (A Scope MD); Department of Dermatology, Hospitalier Lyon Sud, Lyon, France (Prof L Thomas PhD); Lyon Cancer Research Center INSERM U1052—CNRS UMR5286, Lyon, France (Prof L Thomas); Lyon 1 University, Lyon, France (Prof L Thomas); and Dermatology Clinic, Maggiore Hospital, University of Trieste, Trieste, Italy (I Zalaudek MD)

Summary

Background—Whether machine-learning algorithms can diagnose all pigmented skin lesions as accurately as human experts is unclear. The aim of this study was to compare the diagnostic accuracy of state-of-the-art machine-learning algorithms with human readers for all clinically relevant types of benign and malignant pigmented skin lesions.

Methods—For this open, web-based, international, diagnostic study, human readers were asked to diagnose dermatoscopic images selected randomly in 30-image batches from a test set of 1511 images. The diagnoses from human readers were compared with those of 139 algorithms created by 77 machine-learning labs, who participated in the International Skin Imaging Collaboration 2018 challenge and received a training set of 10 015 images in advance. The ground truth of each lesion fell into one of seven predefined disease categories: intraepithelial carcinoma including actinic keratoses and Bowen's disease; basal cell carcinoma; benign keratinocytic lesions including solar lentigo, seborrheic keratosis and lichen planus-like keratosis; dermatofibroma; melanoma; melanocytic nevus; and vascular lesions. The two main outcomes were the differences in the number of correct specific diagnoses per batch between all human readers and the top three algorithms, and between human experts and the top three algorithms.

Findings—Between Aug 4, 2018, and Sept 30, 2018, 511 human readers from 63 countries had at least one attempt in the reader study. 283 (55·4%) of 511 human readers were board-certified dermatologists, 118 (23·1%) were dermatology residents, and 83 (16·2%) were general practitioners. When comparing all human readers with all machine-learning algorithms, the algorithms achieved a mean of 2·01 (95% CI 1·97 to 2·04; p<0·0001) more correct diagnoses (17·91 [SD 3·42] *vs* 19·92 [4·27]). 27 human experts with more than 10 years of experience achieved a mean of 18·78 (SD 3·15) correct answers, compared with 25·43 (1·95) correct answers for the top three machine algorithms (mean difference 6·65, 95% CI 6·06–7·25; p<0·0001). The difference between human experts and the top three algorithms was significantly lower for images in the test set that were collected from sources not included in the training set (human underperformance of 11·4%, 95% CI 9·9–12·9 *vs* 3·6%, 0·8–6·3; p<0·0001).

Interpretation—State-of-the-art machine-learning classifiers outperformed human experts in the diagnosis of pigmented skin lesions and should have a more important role in clinical practice.

However, a possible limitation of these algorithms is their decreased performance for out-of-distribution images, which should be addressed in future research.

Funding—None.

Introduction

Diagnosis of skin cancer needs specific expertise that might not be available in many clinical settings. Accurate diagnosis of early melanoma in particular demands experience in dermatoscopy, a non-invasive examination technique[1] that improves diagnosis compared with examination with the naked eye.[2] Dermatoscopy, which requires proper training and experience, is used widely by dermatologists,[3] but also by general practitioners[4] and other health-care professionals in areas where specialist dermatological services are not readily available.

The paucity of experts and the rising incidence of skin cancer in an aging population[5] have increased the demand for point-of-care decision support systems that can diagnose skin lesions without the need of human expertise. There has been a long tradition of translational research involving machine learning for melanoma diagnosis based on dermatoscopic images.[6–8] Although some automated diagnostic devices have been approved by the US Food and Drug Administration,[9,10] such devices are not widely adopted in clinical practice for various reasons—for example, the devices are approved for melanocytic lesions only and they require preselection of lesions by human experts.

Recent advancements in the field of machine learning, particularly the introduction of convolutional neural networks, have boosted interest in this area of research.[11] Codella and colleagues[12] used ensembles of multiple algorithms to show melanoma recognition accuracies greater than those of expert dermatologists. Subsequently, Esteva and colleagues[13] and Han and colleagues[14] fine-tuned convolutional neural networks with large datasets of clinical images and observed dermatologist-level accuracy for general skin disease classification. Furthermore, Haenssle and colleagues[15] reported expert-level accuracy of algorithms for dermatoscopic images of melanocytic lesions. However, in patients with severe chronic sun damage, up to 50% of pigmented lesions that are biopsied or excised for diagnostic reasons are non-melanocytic.[16]

Training of neural networks for automated diagnosis of pigmented skin lesions has been hampered by the insufficient diversity of available datasets and by selection and verification bias. We tackled this problem by collecting dermatoscopic images of all clinically relevant types of pigmented lesions, and created a publicly available training set of 10 015 images for machine learning.[17] We provided this training set and a test set of 1511 dermatoscopic images to the participants of the International Skin Imaging Collaboration (ISIC) 2018 challenge, with the aim of attracting the best machine-learning labs worldwide to obtain reliable estimates of the accuracy of state-of-the-art machine-learning algorithms. We planned and organised an open, web-based, reader study under the umbrella of the International Dermoscopy Society and invited their members to compare their diagnostic accuracy with that of algorithms. Therefore, the aim of this study was to compare the most

advanced machine-learning algorithms with the most experienced human experts using publicly available data.

Methods

Study design

For this open, web-based, international, diagnostic study, invitations to participate were first issued at the World Congress of Dermoscopy (June 14, 2018) and continued until Sept 28, 2018. 3Gen (San Juan Capistrano, CA, USA) and HealthCert (Singapore) sponsored prizes (a dermatoscope and books) for the best participants. No other compensation was offered to readers. Cumulative numbers of registrations were correlated with specific mailings and social media posts to targeted groups (appendix p 1).

The study protocol was approved by the ethics review boards of the University of Queensland (Brisbane, QLD, Australia) and the Medical University of Vienna (Vienna, Austria), which waived written, informed consent for retrospectively collected and de-identified dermatoscopic images. Before participation, human readers and participants of the ISIC 2018 challenge provided written consent to allow analysis of their ratings.

Procedures

We created a web-based rating platform accessible via username and password on which we ran the screening tests. Upon registration of participants (human readers), we collected information about age, sex, medical education, and years of experience with dermatoscopy. The basic functionality of the platform was to show an image together with a multiple choice question, which included seven predefined disease categories and a single correct answer. Before the main test, each reader had to complete four screening tests, which were used to stratify readers according to skill and to verify if self-reported experience matched actual skill.

The actual survey was done identically to the screening test, but used the test set of 1511 unknown images. The ground truth of each lesion fell into one of seven predefined disease categories: intraepithelial carcinoma including actinic keratoses and Bowen’s disease; basal cell carcinoma; benign keratinocytic lesions including solar lentigo, seborrheic keratosis, and lichen planus-like keratosis; dermatofibroma; melanoma; melanocytic nevus; and vascular lesions. These seven disease categories comprise more than 95% of all pigmented lesions biopsied or excised for diagnostic reasons in clinical practice.[16] As we did not expect human readers to rate all 1511 images, each reader received batches of 30 randomly selected images. Readers could repeat the survey with different batches at their own discretion. Each test set image was rated by a mean of 80 readers (range 43–184; 95% CI 78·6–80·7). We stratified random sampling in four ways to analyse potential effects of class distributions (appendix p 1). The first batch was balanced with regard to number of lesions from each class (balanced), the second batch had more benign lesions (benign; 25 [83%] of 30 lesions), the third more malignant lesions (malignant; 21 [70%] of 30 lesions), and all subsequent batches were randomly drawn from the test set without stratification (random).

We randomly divided a master set of 11 210 dermatoscopic images into a training set (10 015 images; 89·3%) and a test set (1195 images; 10·7%). The images were collected during a period of 20 years from two sites, the Vienna Dermatologic Imaging Research Group (ViDIR) at the Department of Dermatology at the Medical University of Vienna (Vienna, Austria), and the skin cancer practice of Cliff Rosendahl in Queensland (Capalaba, QLD, Australia). The set, which has been described previously,[17] included consecutively collected images of pigmented lesions from different populations. Ground truth was routine pathology evaluation (>50% of all lesions), biology (>1·5 years sequential dermatoscopic imaging without changes), and expert consensus in some cases of common, straightforward, non-melanocytic cases that were not excised. Controversial cases with ambiguous histopathological reports were excluded. The Austrian image set could be divided into the following three subgroups: ViDIR legacy (images captured before 2005 with analog cameras and archived as diapositives), ViDIR current (images captured after 2005 with the DermLite FOTO [3Gen] system or Delta 20 [Heine; Herrsching, Germany], and ViDIR MoleMax (images captured with the MoleMax HD system [Derma Medical Systems; Vienna, Austria]). The Australian image set included lesions from the patients of a primary care facility in an area with high skin cancer incidence. We added 316 images from other centres to the test set (external data), specifically from Turkey, New Zealand, Sweden, and Argentina, to assure diversity of skin types. Our original protocol did not mention test set images from other sources and did not specify the number of disease categories. These amendments were approved by the ethics board of the Medical University of Vienna on Dec 4, 2018.

Predictions of the machine-learning algorithms were provided by the participants of the ISIC 2018 challenge. We co-organised this challenge and an associated workshop[18] at the 21st International Conference On Medical Image Computing & Computer Assisted Intervention, which took place on Sept 20, 2018, in Granada, Spain. Detailed descriptions of submissions can be found at the challenge website. We removed the two lowest scoring (1·4%) of 141 submissions because they produced random predictions because of a formatting error. Machine-learning groups were allowed up to three technically distinct submissions to the challenge, resulting in multiple entries from some groups (there were a total of 139 algorithms from 77 machine-learning labs). For each test case, the class (disease category) with the highest probability was regarded as the diagnosis given by the algorithm.

The two main outcomes were the differences in the number of correct specific diagnoses per batch between human readers and the top three algorithms, and between human experts and the top three algorithms. For a batch of lesions with equal distribution of classes, this difference corresponds to the difference in balanced multiclass accuracy, which is the mean sensitivity calculated for every class in a one-versus-all manner. We chose this metric because it ignores the bias of highly prevalent classes, such as nevi, and gives a good overall estimation of performance in a multiclass setting, as it indirectly measures false positive cases, which are missing in the directly measured true positives of their respective class. Secondary outcomes were differences regarding unbalanced batches.

Statistical analysis

We aimed to include 500 human readers in the study. We used a one-sample *t* test to compare human readers and algorithms and determine whether the difference in the number of correct diagnoses in batches of 30 cases was different from 0. With an SD of 15%, the study had a power of 80% to detect a difference of 1·9% in the number of correct diagnoses at α=0·05.

Because the random batch could be attempted more than once, only the first attempt was included in the analyses of two main outcomes to avoid bias. We calculated the probability of a correct diagnosis for human readers and algorithms by summing the instances of correct diagnoses per lesion and dividing this by the number of readers or number of algorithms.

The probability of correct predictions per lesion, diagnostic values, and area under receiver operator characteristics curves were post-hoc exploratory analyses. For diagnostic values and confusion matrices, we used the majority vote of all ratings for each image. We calculated binary diagnostic values, such as sensitivity and specificity, in a one-versus-all manner. Receiver operating characteristic curves, areas under the curves, and their 95% CIs were calculated with pROC,[19] and we compared areas under the curves with the method described by Delong and colleagues.[20]

Baseline characteristics are reported as n (%) or mean and 95% CI. All p values are two-sided, and p<0·05 was regarded as significant. Bonferroni correction was used for all p values unless otherwise stated. Calculations and plotting were done with R version 3.4.0.[21]

Role of the funding source

There was no funding source for this study. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Between Aug 4, 2018, and Sept 30, 2018, 951 (52·7%) of 1804 potential readers registered on the study platform finished all screening tests, and 511 (28·3%) readers from 63 countries had at least one attempt in the reader study (figure 1). 283 (55·4%) of 511 human readers were board-certified dermatologists, 118 (23·1%) were dermatology residents, and 83 (16·2%) were general practitioners. The distribution of professions in participants of the reader study was similar to users who finished screening, but who did not participate (appendix p 1).

236 (46·2%) of 511 human readers were aged between 31 and 40 years and 321 (62·8%) were female. As the number of years of experience was the most important predictor of a high score in the screening tests, human readers with more than 10 years of experience were regarded as experts.

Human readers achieved a mean of 17·91 (SD 3·42) correct answers (of a possible 30) in the balanced batch, compared with 25·85 (1·83) correct answers for the top three machine-learning algorithms (MetaOptima Technology Inc, DAISYLab, and Medical Image Analysis

Group, Sun Yat-sen University; mean difference 7·94, 95% CI 7·76–8·12; p<0·0001). 27 human experts with more than 10 years of experience achieved a mean of 18·78 (SD 3·15) correct answers, compared with 25·43 (1·95) correct answers for the top three machine-learning algorithms (mean difference 6·65, 95% CI 6·06–7·25; p<0·0001; figure 2). We found similar results for the top three human experts compared with the top three algorithms (appendix p 3). The mean difference between experts and the top three algorithms was smaller in benign, malignant, and random batches (5·39, 95% CI 4·64–6·15; 5·20, 4·21–6·18; and 3·76, 3·08–4·43, respectively; figure 2) compared with the balanced batch.

When comparing all human readers with all machine-learning algorithms, the algorithms achieved a mean of 2·01 (95% CI 1·97 to 2·04; p<0·0001) more correct diagnoses (17·91 [SD 3·42] *vs* 19·92 [4·27]; appendix p 4). All algorithms had a mean 0·79 (95% CI 0·64 to 0·94; p<0·0001) more correct diagnoses than did expert readers (figure 3). The difference between human readers and algorithms was greater in batches with more benign cases (mean difference 1·59, 95% CI 1·44 to 1·74; p<0·0001) and in random batches (mean difference 1·06, 0·93 to 1·20; p<0·0001). In malignant batches, human experts outperformed algorithms (mean difference −0·54, −0·76 to −0·33; p<0·0001).

The probability for correct diagnosis of an image increased with the number of years of experience of the human reader and depended on the image source. For experts, the highest probability of a correct diagnosis was found in the ViDIR MoleMax dataset (91·4%, 95% CI 90·1–92·7) and the lowest in the Australian dataset (60·1%, 56·0–64·1; appendix p 5). Compared with other image sets, the difference between experts and the top three algorithms was significantly lower for images that were collected from centres that did not provide images to the training set (human underperformance of 11·4%, 95% CI 9·9–12·9 *vs* 3·6%, 0·8–6·3; p<0·0001).

The mean sensitivity across all classes was 79·2% (95% CI 64·4–94·0) for all human readers, 81·2% (66·1–96·3) for experts, and 88·5% (82·2–94·7; MetaOptima Technology Inc), 85·6% (79·1–92·0; DAISYLab), and 84·5% (78·5–90·5; Medical Image Analysis Group, Sun Yat-sen University) for the top three algorithms. The sensitivity for most malignant classes (melanoma, actinic keratosis, and Bowen's disease) was higher for the top three algorithms than for experts (table; appendix p 7). We observed the largest difference between human experts and algorithms with regard to the sensitivity for intraepithelial carcinoma (51·2%, 95% CI 35·5–66·7 *vs* 90·7%, 77·9–97·4; table), which were commonly misdiagnosed by human readers, whereas the errors of algorithms were more evenly distributed across classes (appendix p 6).

The point indicating the mean sensitivity (0·76, 95% CI 0·74–0·77) and specificity (0·78, 0·77–0·79) of human readers was situated below the receiver operating characteristic curves of the top three algorithms (figure 4). The area under the curve for the prediction of malignancy via vote frequency was 0·958 (95% CI 0·948–0·967) for human readers, and 0·963 (0·953–0·973; p=0·46; MetaOptima Technology Inc), 0·971 (0·961–0·982; p=0·05; DAISYLab), and 0·958 (0·945–0·972; p=0·91; Medical Image Analysis Group, Sun Yat-Sen University) for the top three algorithms (no significant difference for all three comparisons).

Discussion

We provide a state-of-the-art comparison of machine-learning algorithms with human readers for the diagnosis of all clinically relevant types of pigmented skin lesions using dermatoscopic images. Machine-learning algorithms outperformed human readers with respect to most outcome measures. In sets of 30 randomly selected lesions, the best machine-learning algorithms achieved a mean of 7·94 more correct diagnoses than the average human reader, and a mean of 6·65 more correct diagnoses than expert readers.

A common problem in human reader studies is the definition of experts. In a screening test, we compared the self-reported domain-specific experience of participants with their actual performance and found that self-reported years of experience reliably predicted domain-specific expertise (appendix p 2). Unlike in similar studies,[15,22,23] our test set included not only melanoma and nevi, but also non-melanocytic lesions. The primary task in our study was a multiclass problem with seven disease categories, and not just the simple binary problem of melanoma versus nevi. Therefore, our diagnostic study could be considered closer to a real-life situation than other studies in this field. Our test set is unique because of the large number of benign lesions that were not biopsied or excised. Inclusion of typical benign lesions avoids verification bias, which is a common limitation of diagnostic studies. Most benign lesions were nevi that we monitored for more than 18 months without any changes, which is as reliable a ground truth as pathological verification. The lesions were collected in two different settings—a tertiary referral centre in Europe and a skin cancer clinic in Australia. European patients are typified by a high number of nevi and a personal history of melanoma, and Australian patients by severe chronic sun damage. Human readers, including experts, achieved the lowest accuracy in the Australian dataset, which is not surprising since this dataset was more challenging and contained many equivocal lesions on chronic sun damaged skin that were biopsied to rule out malignancy. This set also contained difficult to diagnose melanomas and many pigmented intraepithelial carcinomas, which were often misdiagnosed by human readers. However, the top three algorithms performed equally well across all datasets, including the Australian set, and across all diagnoses, including pigmented intraepithelial carcinomas.

Overfitting to the distribution of images in the training set might explain the superior performance of algorithms. However, overfitting would lead to lack of generalisability. We anticipated overfitting and tried to quantify it by including a set of images from sources that did not provide images for the training set. As we expected, the accuracy of the top three machine-learning algorithms was lower in the set of new lesions, but still higher than the accuracy of human experts, which was also shown previously by Han and colleagues.[14] This result indicates a potential limitation of algorithms for out-of-distribution images, which should be addressed in future research.

The low sensitivity of human experts for melanoma is striking and might be explained by the difficult test set, especially with regard to the Australian set, and by the framing of the task and presentation of images. A limitation of our study is that we did not provide additional data, for example, anatomical site, age, and sex, beyond dermatoscopic images, although these data were also lacking in the development of the algorithm. In a real-world situation,

human readers would consider the variability of lesions within a given patient. This approach, which is a variant of the so-called ugly duckling rule,[24] increases sensitivity and specificity, but requires examination of the entire patient and not just single lesions. Therefore, our diagnostic study deviated from a real-world scenario and simulates a telemedical approach, which could be a future domain for machine-learning algorithms.

Another obstacle for human readers was that the lesions in the test set and training set were not standardised. The images were photographed with different devices and magnifications but, in reality, human readers could be used to a single device with fixed magnification and constant representation of colours. However, the variations in the dataset are representative of the variations observed in the field of skin imaging, which are a consequence of the high diversity of dermatoscopes and cameras, and the absence of applied standards.[25] We asked human readers to rate lesions from the training set to get used to the diversity of the test set to mitigate this effect.

Although machine-learning algorithms outperformed human experts in nearly every aspect, higher accuracy in a diagnostic study with digital images does not necessarily mean better clinical performance or patient management.[26] The metrics used in this study treated all diagnoses equally. The algorithms were trained to optimise the mean sensitivity across all classes, and did not consider that it is more detrimental to mistake a malignant for a benign lesion than vice versa. We deliberately chose a balanced metric because the test set was highly imbalanced towards nevi, and we wanted to penalise strategies that optimise accuracy by preferring predictions in favour of the most prevalent class. However, in practice, diagnosis of a melanoma as a basal cell carcinoma will be of no major clinical consequence for a patient with regard to primary diagnostic tests, because both lesions are usually excised or biopsied. Therefore, a metric that is based on the binary outcome of benign or malignant (or excise or dismiss) might be more clinically relevant. When we dichotomised the diagnostic classes into a benign and a malignant group and compared the accuracy of the majority vote of human readers with the top three algorithms, we found no difference in the area under the curve. Similar findings were reported in radiology, where so-called swarm intelligence improved the diagnostic accuracy of human readers.[27]

Although the lack of superiority in melanoma sensitivity of experts compared with the average human reader was outweighed by the superiority of experts for other diagnoses, this fact deserves an explanation. We hypothesise that, given their lower level of confidence, the non-expert readers tended to give false positives for melanoma, since the cost of a false negative decision on a possible melanoma is more severe than the cost of a false positive. The expert readers, who had a higher level of confidence, preferred to use their highest likelihood prediction.

Our study is a simulation and deviates from a real-life setting. In a real-life setting, evaluation of skin lesions is not limited to a timeframe of 20 s and human readers might make different decisions when faced with a patient in person. In future, it is probable that automated classifiers will be used under human guidance, rather than alone.[28] Hence, it might be more appropriate to test the accuracy of automated classifying algorithms in the hand of human readers rather than to test classifiers and humans alone.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the International Dermoscopy Society for administrative support and for contacting members on their mailing lists. We thank HealthCert for providing five books, and 3Gen for providing a dermatoscope as prize for study participants. AM, MM, and AH are supported by a National Institutes of Health support grant to Memorial Sloan Kettering Cancer Center (P30 CA008748). We express our gratitude for our community partners and the machine-learning engineers and human readers who participated in the study.

Declaration of interests

PT reports personal fees from Silverchair and grants from MetaOptima Technology, outside the submitted work. NC reports salary from IBM as an employee during the conduct of the study, and capital gains and dividend income from a diversified portfolio in tech and health-care industries outside the submitted work. NC has a patent (multiple IBM owned patent applications in the space of skin assessment and analysis) pending. HK reports non-financial support from Derma Medical Systems, non-financial support from Fotofinder, personal fees from Almirall, personal fees from Health Cert, and personal fees from PelPharma, all outside the submitted work. JM reports grants and personal fees from Amgen, grants, personal fees, and non-financial support from Almirall, personal fees from Pierre Fabre, personal fees and trial conduction for Bristol-Myers Squibb, grants from GlaxoSmithKline, personal fees and non-financial support from Isdin, personal fees and non-financial support from Canfield, grants from Novartis, grants and trial conduction from Merck Sharp & Dohme, grants and non-financial support from La Roche Posay, non-financial support from Mavig, non-financial support from 3Gen, personal fees and potential spouse's conflict of interest for Sanofi, trial conduction with Pfizer, grants, personal fees, non-financial support, and trial conduction with Scibase, potential spouse's conflict of interest from Ojer Pharma, grants, personal fees, trial conduction, and potential spouse's conflict of interest with Roche, and personal fees from Sun Pharma, all outside the submitted work. AO reports personal fees from MoleMap New Zealand and personal fees from DermNet New Zealand, outside the submitted work. SP reports grants and personal fees for her spouse from Amgen, grants, personal fees, and non-financial support from Almirall, personal fees from Pierre Fabre, personal fees and trial conduction from Bristol-Myers Squibb, grants from GlaxoSmithKline, personal fees and non-financial support from Isdin, personal fees and non-financial support for her spouse from Canfield, grants from Novartis, grants from and conduction of trials with Merck Sharp & Dohme, grants and non-financial support from La Roche Posay, non-financial support from Mavig, non-financial support from 3Gen, personal fees and trial conduction from Sanofi, trial conduction from Pfizer, grants, non-financial support, trial conduction, and potential spousal conflict of interest from Scibase, personal fees from Ojer Pharma, grants, personal fees, and trial conduction from Roche, and personal fees for her spouse from Sun Pharma, all outside the submitted work. HPS reports grants and personal fees from Canfield Scientific, personal fees and role as medical advisory board member, medical consultant, and minor shareholder from MoleMap NZ, personal fees and role as medical consultant and shareholder from e-dermconsult, personal fees from MetaOptima, and grants from Fotofinder, all outside the submitted work. All other authors declare no competing interests.

References

1. Saphier J Die Dermatoskopie. Arch f Dermat 1921; 128: 1–19.
2. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. Lancet Oncol 2002; 3: 159–65. [PubMed: 11902502]
3. Forsea AM, Tschandl P, Del Marmol V, et al. Factors driving the use of dermoscopy in Europe: a pan-European survey. Br J Dermatol 2016; 175: 1329–37. [PubMed: 27469990]
4. Rosendahl C, Williams G, Eley D, et al. The impact of subspecialization and dermatoscopy use on accuracy of melanoma diagnosis among primary care doctors in Australia. J Am Acad Dermatol 2012; 67: 846–52. [PubMed: 22325462]
5. Rogers HW, Weinstock MA, Feldman SR, Coldiron BM. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012. JAMA Dermatol 2015; 151: 1081–86. [PubMed: 25928283]
6. Binder M, Steiner A, Schwarz M, Knollmayer S, Wolff K, Pehamberger H. Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study. Br J Dermatol 1994; 130: 460–65. [PubMed: 8186110]

7. Menzies SW, Bischof L, Talbot H, et al. The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. Arch Dermatol 2005; 141: 1388–96. [PubMed: 16301386]
8. Dreiseitl S, Binder M, Hable K, Kittler H. Computer versus human diagnosis of melanoma: evaluation of the feasibility of an automated diagnostic system in a prospective clinical trial. Melanoma Res 2009; 19: 180–84. [PubMed: 19369900]
9. Monheit G, Cognetta AB, Ferris L, et al. The performance of MelaFind: a prospective multicenter study. Arch Dermatol 2011; 147: 188–94. [PubMed: 20956633]
10. Malvehy J, Hauschild A, Curiel-Lewandrowski C, et al. Clinical performance of the Nevisense system in cutaneous melanoma detection: an international, multicentre, prospective and blinded clinical trial on efficacy and safety. Br J Dermatol 2014; 171: 1099–107. [PubMed: 24841846]
11. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998; 86: 2278–324.
12. Codella N, Nguyen Q-B, Pankanti S, et al. Deep learning ensembles for melanoma recognition in dermoscopy images. IBM J Res Dev 2017; 61: 1–15. [PubMed: 29200477]
13. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115–18. [PubMed: 28117445]
14. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 2018; 138: 1529–38. [PubMed: 29428356]
15. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018; 29: 1836–42. [PubMed: 29846502]
16. Rosendahl C, Tschandl P, Cameron A, Kittler H. Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions. J Am Acad Dermatol 2011; 64: 1068–73. [PubMed: 21440329]
17. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018; 5: 180161. [PubMed: 30106392]
18. Stoyanov D, Taylor Z, Sarikaya D, et al. OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis. First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings. New York: Springer, 2018.
19. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12: 77. [PubMed: 21414208]
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44: 837–45. [PubMed: 3203132]
21. Wickham H ggplot2: elegant graphics for data analysis. New York: Springer, 2016.
22. Codella NCF, Gutman D, Emre Celebi M, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). arXiv 2017; published online Oct 17. DOI:1710.05006.
23. Marchetti MA, Codella NCF, Dusza SW, et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018; 78: 270–77. [PubMed: 28969863]
24. Gaudy-Marqueste C, Wazaefi Y, Bruneu Y, et al. Ugly duckling sign as a major factor of efficiency in melanoma detection. JAMA Dermatol 2017; 153: 279–84. [PubMed: 28196213]
25. Finnane A, Curiel-Lewandrowski C, Wimberley G, et al. Proposed technical guidelines for the acquisition of clinical images of skin-related conditions. JAMA Dermatol 2017; 153: 453–57. [PubMed: 28241182]

26. Cook DA, Sherbino J, Durning SJ. Management reasoning: beyond the diagnosis. JAMA 2018; 319: 2267–68. [PubMed: 29800012]

27. Rosenberg L, Lungren M, Halabi S, Willcox G, Baltaxe D, Lyons MM. Artificial swarm intelligence employed to amplify diagnostic accuracy in radiology. IEE IEMCON 2018; Vancouver; 11 1–3, 2018.

28. Codella NCF, Lin C-C, Halpern A, Hind M, Feris R, Smith JR. Collaborative human-AI (CHAI): Evidence-based interpretable melanoma classification in dermoscopic images. Understanding and interpreting machine learning in medical image computing applications. New York: Springer International Publishing, 2018.

Research in context

Evidence before this study

We searched the online databases Medline, arXiv, and PubMed Central using the search terms “melanoma diagnosis” or “melanoma detection” for articles published between Jan 1, 2002, and Dec 15, 2017, in English. After screening 1375 abstracts, we found 90 studies that investigated the accuracy of automated diagnostic systems for the diagnosis of melanoma. 57 studies provided enough data for a quantitative analysis and nine made direct comparisons with human experts. The summary estimate of the accuracy of machine-learning algorithms was on par with, but did not exceed, human experts. Many studies did not use an independent, external test set and we found no study that fully covered the heterogeneity of pigmented lesions by including all relevant types of non-melanocytic lesions. Many studies were also prone to different types of biases, including selection and verification bias, and did not use publicly available data. Most studies focused on a single machine-learning algorithm and compared it with a small number (less than 100) of human readers.

Added value of this study

We provide a state-of-the-art comparison of the most advanced machine-learning algorithms with a large number of human readers, including the most experienced human experts. We included all types of clinically relevant pigmented skin lesions, not only melanoma and nevi, and algorithms and humans were tested with publicly available images, including images from sites with different populations and skin types. Most algorithms were also trained with a standard image set; hence, performance should be easily reproducible by other research teams. Our results show that state-of-the-art machine-learning algorithms outperform even the most experienced human experts.

Implications of all the available evidence

The results of our study could improve the accuracy of the diagnosis of pigmented skin lesions in areas where specialist dermatological service is not readily available, and might accelerate the acceptance and implementation of automated diagnostic devices in the field of skin cancer diagnosis.

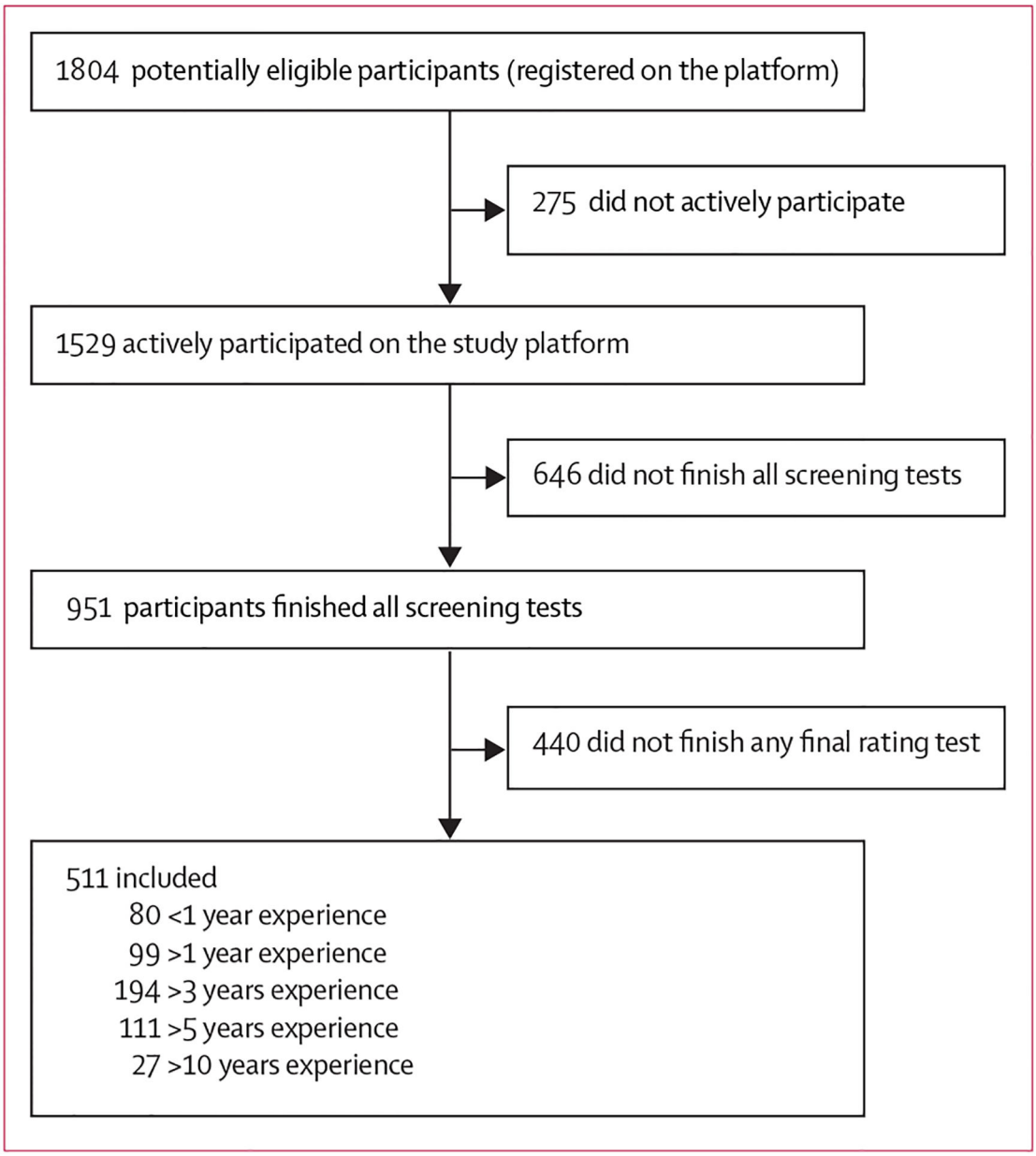


Figure 1:
Numbers of registered and participating users on the study platform

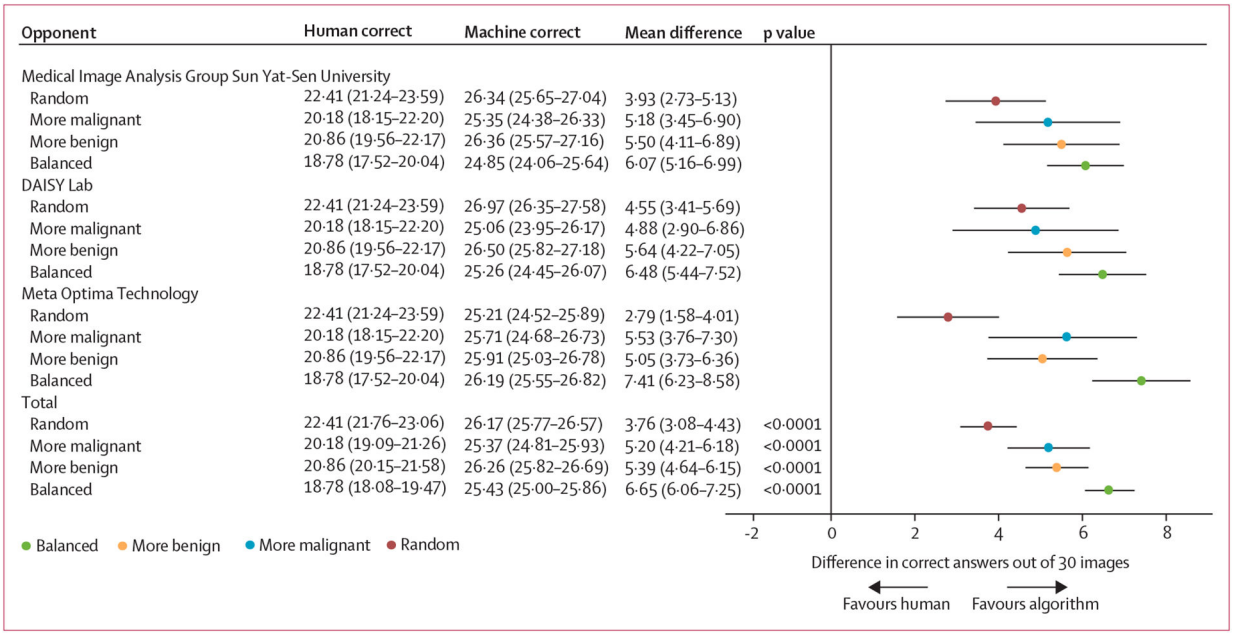
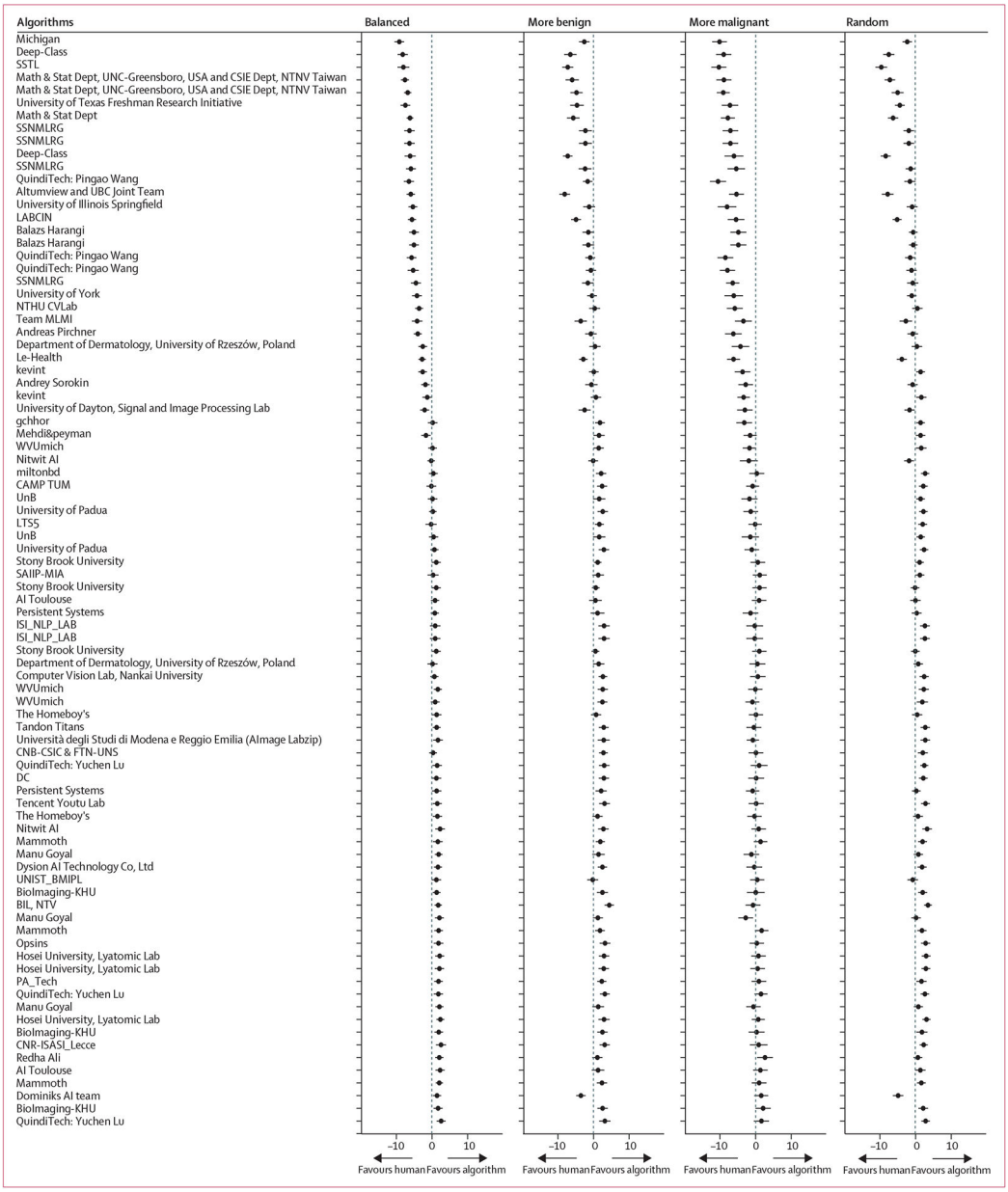


Figure 2: Mean differences in correct diagnoses of human experts versus the top three machine-learning algorithms in batches of 30 images

Data are mean (95% CI).



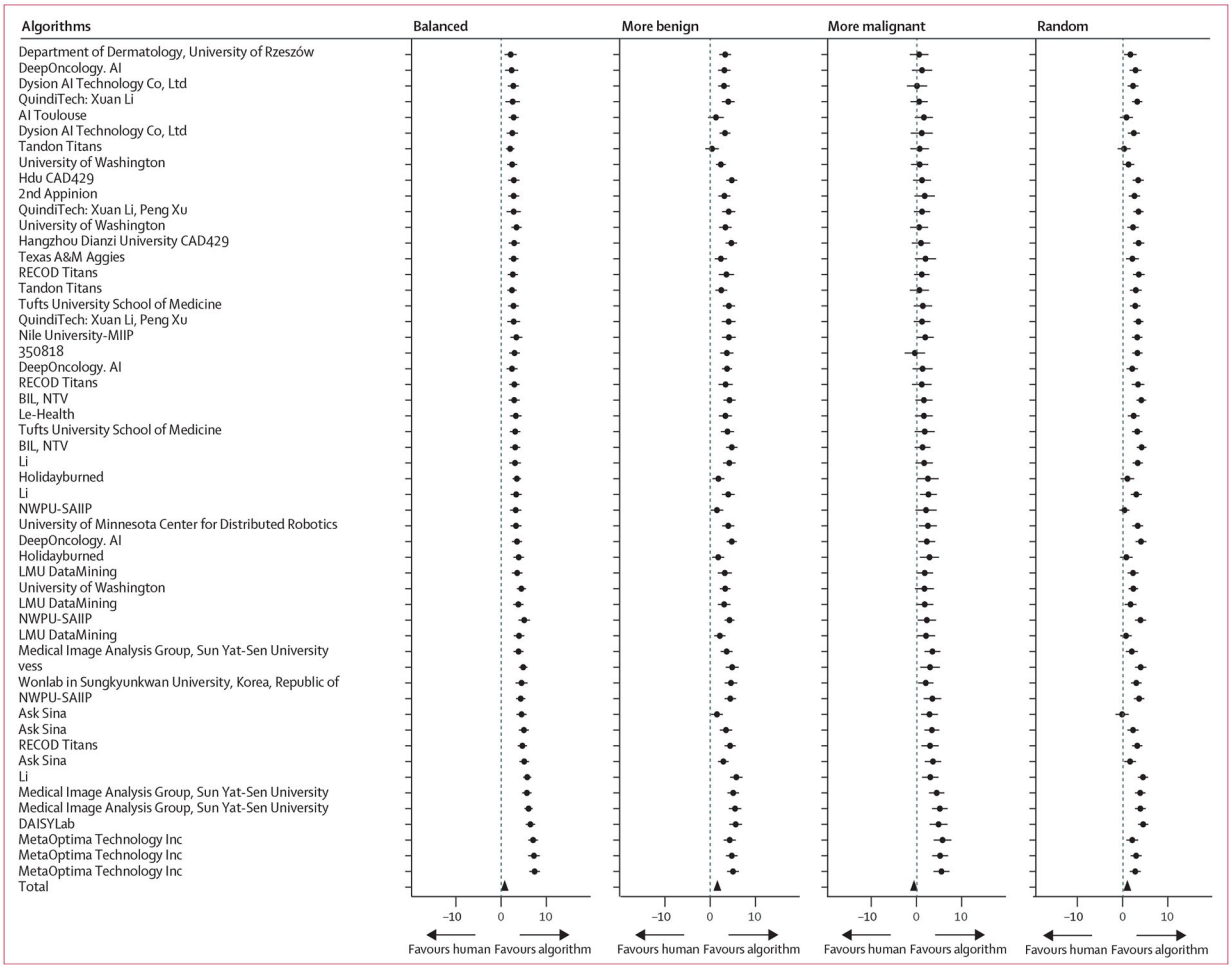


Figure 3: Mean difference between all human expert readers and all machine-learning algorithms for the number of correct diagnoses per batch

Error bars denote 95% CIs. Machine-learning groups were allowed up to three technically distinct test set submissions resulting in multiple entries for some groups. The performance of each algorithm *vs* humans is increased the further down the y axis they are listed.

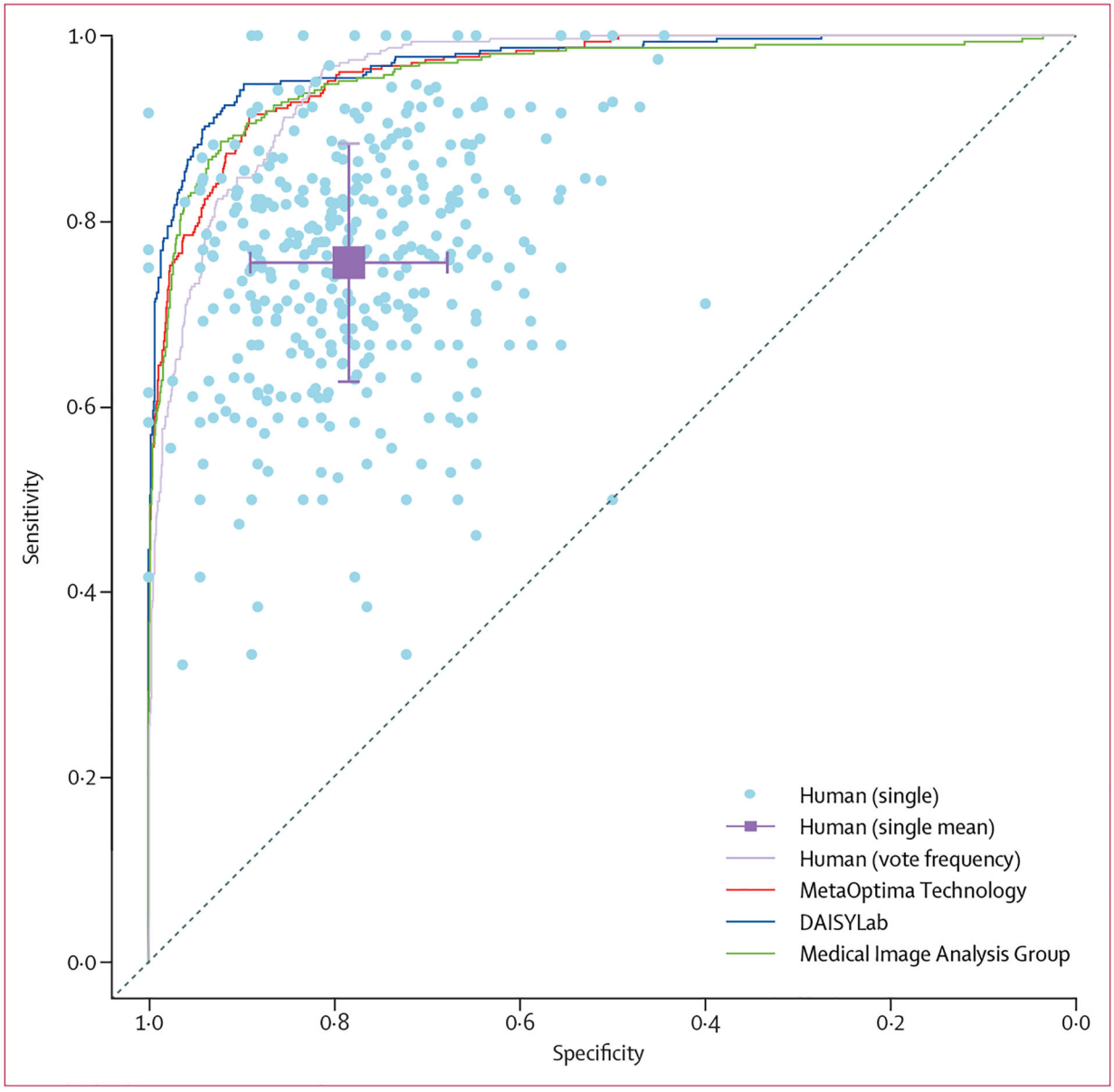


Figure 4: Receiver operating characteristic curves of the diagnostic performance for discrimination of malignant from benign pigmented skin lesions

Blue dots indicate single human sensitivities and specificities, the purple box indicates the mean, and the error bars around the mean indicate 95% CI.

Table:

Diagnostic performance of all human readers, expert readers, all machine learning algorithms, and top three algorithms per diagnostic category

	Sensitivity	Specificity	Negative predictive value	Positive predictive value
All human readers (n=511)				
Actinic keratoses and Bowen's disease	48·8% (33·3–64·5)	98·4% (97·7–99·0)	98·5% (97·7–99·1)	47·7% (32·5–63·3)
Basal cell carcinoma	86·0% (77·3–92·3)	98·9% (98·2–99·4)	99·1% (98·4–99·5)	83·3% (74·4–90·2)
Benign keratinocytic lesions	80·6% (74·8–85·7)	95·4% (94·1–96·4)	96·7% (95·6–97·6)	74·5% (68·4–79·9)
Dermatofibroma	77·3% (62·2–88·5)	99·6% (99·1–99·8)	99·3% (98·8–99·7)	85·0% (70·2–94·3)
Melanoma	73·1% (65·8–79·6)	92·8% (91·3–94·2)	96·4% (95·3–97·4)	56·6% (49·7–63·2)
Melanocytic nevus	88·8% (86·5–90·7)	95·2% (93·2–96·8)	84·9% (82·0–87·5)	96·5% (95·1–97·7)
Vascular lesions	100·0% (90·0–100·0)	99·7% (99·2–99·9)	100·0% (99·7–100·0)	87·5% (73·2–95·8)
Expert readers (n=27)				
Actinic keratoses and Bowen's disease	51·2% (35·5–66·7)	98·6% (97·8–99·1)	98·6% (97·8–99·1)	51·2% (35·5–66·7)
Basal cell carcinoma	89·2% (81·1–94·7)	99·1% (98·4–99·5)	99·3% (98·7–99·7)	86·5% (78·0–92·6)
Benign keratinocytic lesions	84·3% (78·8–88·9)	95·8% (94·6–96·8)	97·3% (96·3–98·1)	77·2% (71·3–82·4)
Dermatofibroma	86·4% (72·6–94·8)	99·1% (98·5–99·5)	99·6% (99·1–99·8)	74·5% (60·4–85·7)
Melanoma	67·8% (60·3–74·8)	94·0% (92·5–95·2)	95·8% (94·6–96·8)	58·9% (51·7–65·8)
Melanocytic nevus	89·3% (87·1–91·3)	94·2% (92·0–95·9)	85·4% (82·5–88·0)	95·9% (94·3–97·1)
Vascular lesions	100·0% (90·0–100·0)	99·6% (99·1–99·9)	100·0% (99·7–100·0)	85·4% (70·8–94·4)
All algorithms (n=139)				
Actinic keratoses and Bowen's disease	68·2% (52·4–81·4)	99·4% (98·8–99·7)	99·1% (98·4–99·5)	76·9% (60·7–88·9)
Basal cell carcinoma	84·9% (76·0–91·5)	98·4% (97·6–99·0)	99·0% (98·3–99·5)	77·5% (68·1–85·1)
Benign keratinocytic lesions	78·1% (72·0–83·4)	98·4% (97·5–99·0)	96·4% (95·2–97·3)	89·1% (83·8–93·1)
Dermatofibroma	68·9% (53·4–81·8)	100·0% (99·7–100·0)	99·1% (98·4–99·5)	100·0% (88·8–100·0)
Melanoma	67·3% (59·7–74·2)	97·0% (96·0–97·9)	95·9% (94·7–96·9)	74·2% (66·6–80·9)
Melanocytic nevus	96·3% (94·8–97·4)	84·7% (81·6–87·5)	93·8% (91·4–95·7)	90·4% (88·4–92·2)
Vascular lesions	77·1% (59·9–89·6)	99·9% (99·5–100·0)	99·5% (98·9–99·8)	93·1% (77·2–99·2)
Top three algorithms (n=3)				
Actinic keratoses and Bowen's disease	90·7% (77·9–97·4)	98·5% (97·7–99·0)	99·7% (99·3–99·9)	62·9% (49·7–74·8)
Basal cell carcinoma	88·4% (80·2–94·1)	98·4% (97·6–99·0)	99·2% (98·6–99·6)	78·5% (69·5–85·9)
Benign keratinocytic lesions	83·8% (78·4–88·4)	98·3% (97·4–98·9)	97·2% (96·2–98·1)	89·3% (84·4–93·1)

	Sensitivity	Specificity	Negative predictive value	Positive predictive value
Dermatofibroma	81·8% (67·3–91·8)	99·3% (98·8–99·7)	99·5% (99·0–99·8)	78·3% (63·6–89·1)
Melanoma	81·9% (75·4–87·3)	96·2% (95·1–97·2)	97·6% (96·7–98·4)	73·6% (66·9–79·6)
Melanocytic nevus	91·6% (89·7–93·3)	94·2% (92·1–95·9)	88·3% (85·6–90·6)	96·0% (94·4–97·2)
Vascular lesions	89·2% (74·6–97·0)	99·5% (99·1–99·8)	99·7% (99·3–99·9)	82·5% (67·2–92·7)

Data are mean (95% CI). Majority vote was used to calculate the sensitivities, specificities, and positive and negative predictive values per group.